Iniziato	martedì, 7 gennaio 2020, 15:07
Stato	Completato
Terminato	martedì, 7 gennaio 2020, 15:36
Tempo impiegato	28 min. 43 secondi
Punteggio	10,50/15,00
Valutazione	**21,00** su un massimo di 30,00 (**70**%)

Domanda **1**

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

In order to reduce the dimensionality of a dataset, which is the advantage of Multi Dimensional Scaling (MDS), with respect to Principal Component Analysis (PCA)

Scegli un'alternativa:

- a. MDS can be used also with categorical data, provided that the matrix of the distance is available, while PCA is limited to vector spaces ✔
- b. MDS can be used with categorical data after a transformation in a vector space
- c. MDS requires less computational power
- d. MDS can work on any kind of data, while PCA is limited to categorical data

Risposta corretta.

La risposta corretta è: MDS can be used also with categorical data, provided that the matrix of the distance is available, while PCA is limited to vector spaces

Domanda **2**

Parzialmente corretta

Punteggio ottenuto 0,50 su 1,00

Which is the purpose of discretisation?

Scegli un'alternativa:

- a. Increase the number of distinct values in an attribute, in order to put in evidence possible patterns and regularities
- b. Reduce the range of values of a numeric attribute, to make all the attributes more comparable
- c. Reduce the number of distinct values in an attribute, in order to put in evidence possible patterns and regularities
- d. Reduce the number of distinct values in an attribute, in order to increase the efficiency of the computation ☑ The increased efficiency of the computations can be a byproduct, but the main reason is to put in evidence possible patterns and regularities

Risposta parzialmente esatta.

La risposta corretta è: Reduce the number of distinct values in an attribute, in order to put in evidence possible patterns and regularities

Domanda **3**

Risposta errata

Punteggio ottenuto 0,00 su 1,00

Given the two binary vectors below, which is their similarity according to the Simple Matching Coefficient?

a b c d e f g h i j
1 0 0 0 1 0 1 1 0 1
1 0 1 1 1 0 1 0 1 0

Scegli un'alternativa:

- a. 0.1
- b. 0.5
- c. 0.2
- d. 0.3 ✖ No, this is only the fraction of matching 1's, which is computed by the Jaccard coefficient

Risposta errata.

La risposta corretta è: 0.5

Domanda **4**

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

Which of the following statements regarding the discovery of association rules is true? (One or more)

Scegli una o più alternative:

- a. The support of a rule can be computed given the confidence of the rule
- b. The support of an itemset is anti-monotonic with respect to the composition of the itemset ✔
- c. The confidence of an itemset is anti-monotonic with respect to the composition of the itemset
- d. The confidence of a rule can be computed starting from the supports of itemsets ✔

Your answer is correct.

Le risposte corrette sono: The confidence of a rule can be computed starting from the supports of itemsets, The support of an itemset is anti-monotonic with respect to the composition of the itemset

Domanda **5**

Risposta errata

Punteggio ottenuto 0,00 su 1,00

Given the definitions below:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

which of the formulas below computes the *precision* of a binary classifier?

Scegli un'alternativa:

- a. (TP + TN) / (TP + FP + TN + FN)
- b. TN / (TN + FP)
- c. TP / (TP + FP)
- d. TP / (TP + FN) ✖ No, this is called *sensitivity*, or *hit rate* or *recall*, which is the number of detected true positives divided by the total number of positives

Risposta errata.

La risposta corretta è: TP / (TP + FP)

Domanda **6**

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

A Decision Tree is...

Scegli un'alternativa:

- a. A tree-structured plan of tests on multiple attributes to forecast the target
- b. A tree-structured plan of tests on single attributes to obtain the maximum purity of a node
- c. A tree-structured plan of tests on single attributes to forecast the cluster
- d. A tree-structured plan of tests on single attributes to forecast the target ✔

Risposta corretta.

La risposta corretta è: A tree-structured plan of tests on single attributes to forecast the target

Domanda **7**

Risposta errata

Punteggio ottenuto 0,00 su 1,00

Given the two binary vectors below, which is their similarity according to the Jaccard Coefficient?

a b c d e f g h i j

1 0 0 0 1 0 1 1 0 1

1 0 1 1 1 0 1 0 1 0

Scegli un'alternativa:

- a. 0.375
- b. 0.5 ✖ No, this is the fraction of all the matches, that is computed by the Simple Matching Coefficient
- c. 0.1
- d. 0.2

Risposta errata.

It is the number of matching 1 divided by the number of matching 1 + the number of non-matching

La risposta corretta è: 0.375

Domanda **8**

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

Which is different from the others?

Scegli un'alternativa:

- a. K-means
- b. Decision Tree ✔ This is not a clustering method
- c. Dbscan
- d. Expectation Maximisation

Risposta corretta.

La risposta corretta è: Decision Tree

Domanda **9**

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

After fitting DBSCAN with the default parameter values the results are: 0 clusters, 100% of noise points. Which will be your next trial?

Scegli una o più alternative:

- a. Reduce the minimum number of objects in the neighborhood ✔
- b. Decrease the radius of the neighborhood
- c. Reduce the minimum number of objects in the neighborhood and the radius of the neighborhood
- d. Increase the radius of the neighborhood ✔

Risposta corretta.

Le risposte corrette sono: Reduce the minimum number of objects in the neighborhood, Increase the radius of the neighborhood

Domanda **10**

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

How does *pruning* work when generating frequent itemsets?

Scegli un'alternativa:

- a. If an itemset is frequent, then none of its supersets can be frequent, therefore the frequencies of the supersets are not evaluated
- b. If an itemset is not frequent, then none of its subsets can be frequent, therefore the frequencies of the subsets are not evaluated
- c. If an itemset is frequent, then none of its subsets can be frequent, therefore the frequencies of the subsets are not evaluated
- d. If an itemset is not frequent, then none of its supersets can be frequent, therefore the frequencies of the supersets are not evaluated ✔

Risposta corretta.

La risposta corretta è: If an itemset is not frequent, then none of its supersets can be frequent, therefore the frequencies of the supersets are not evaluated

Domanda **11**

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

What measure is maximised by the Expectation Masimisation algirithm for clustering?

Scegli un'alternativa:

- a. The support of a class
- b. The *likelihood* of a class label, given the values of the attributes of the example ✔
- c. The likelihood of an example
- d. The likelihood of an attribute, given the class label

Your answer is correct.

La risposta corretta è: The *likelihood* of a class label, given the values of the attributes of the example

Domanda **12**

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

Which of the following preprocessing activities is useful to build a Naive Bayes classifier if the independence hypothesis is violated

Scegli un'alternativa:

- a. Feature selection ✔
- b. Standardisation
- c. Normalisation
- d. Discretisation

Risposta corretta.

La risposta corretta è: Feature selection

Domanda **13**

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

What does K-means try to minimise?

Scegli un'alternativa:

- a. The *separation*, that is the sum of the squared distances of each point with respect to its centroid
- b. The *separation*, that is the sum of the squared distances of each cluster centroid with respect tho the global centroid of the dataset
- c. The *distortion*, that is the sum of the squared distances of each point with respect to the points of the other clusters
- d. The *distortion*, that is the sum of the squared distances of each point with respect to its centroid ✔

Risposta corretta.

La risposta corretta è: The *distortion*, that is the sum of the squared distances of each point with respect to its centroid

Domanda **14**

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

What is the *cross validation*

Scegli un'alternativa:

- a. A technique to improve the speed of a classifier
- b. A technique to obtain a good estimation of the performance of a classifier when it will be used with data different from the training set ✔
- c. A technique to improve the quality of a classifier
- d. A technique to obtain a good estimation of the performance of a classifier with the training set

Risposta corretta.

La risposta corretta è: A technique to obtain a good estimation of the performance of a classifier when it will be used with data different from the training set

Domanda **15**

Risposta errata

Punteggio ottenuto 0,00 su 1,00

The *information gain* is used to

Scegli un'alternativa:

- a. select the attribute which maximises, for a given test set, the ability to predict the class value ✖ No, the information gain is computed on the *training set*, not on the test set
- b. select the attribute which maximises, for a given training set, the ability to predict all the other attribute values
- c. select the attribute which maximises, for a given training set, the ability to predict the class value
- d. select the class with maximum probability

Your answer is incorrect.

La risposta corretta è: select the attribute which maximises, for a given training set, the ability to predict the class value

◄ Lab Activity 17-12-2019 - Simulation of lab exa

Vai a...

Introduction to Big Data - Slides ►